11-16-00

A

ASSISTANT COMMISSIONER FOR PATENTS
BOX PATENT APPLICATION
Washington, D.C. 20231

jc945 U.S. PTO
11/15/00

DOCKET NO. YOR920000807US1
Date: November 15, 2000

Sir:

jc829 U.S. PTO
09/713015
11/15/00

under Rule 1.53(f)
Transmitted herewith for filing is the Patent Application/of:

Inventors: Eric William Brown, Anni R. Coden, John Martin Prager, Dragomir Radkov Radev, Valerie Samn

For: SYSTEM AND METHOD FOR FINDING THE MOST LIKELY ANSWER TO A NATURAL LANGUAGE QUESTION

Enclosed are:

X 9 Sheets of Informal Drawings.

____ An assignment of the invention to International Business Machines Corporation, Armonk, New York 10504.

____ A certified copy of a _____ application.

X Declaration and Power of Attorney is attached to the application. (unsigned)

____ Associate Power of Attorney.

____ Information Disclosure Statement with form PTO-1449 with references attached.

The filing fee has been calculated as shown below:

	(Col. 1)	(Col. 2)	OTHER THAN A SMALL ENTITY	
FOR:	NO. FILED	NO. EXTRA	RATE	FEE
BASIC FEE				$ 710.00
TOTAL CLAIMS	12 - 20 =	0	X $ 18 =	$ 0.00
INDEP CLAIMS	3 - 3 =	0	X $ 80 =	$ 0.00
___ MULTIPLE DEPENDENT CLAIM PRESENTED			+ $ 270=	$ 0.00
			TOTAL	$ 710.00

If the difference in Col. 1 is less than zero, enter "0" in Col. 2.

x Please charge my Deposit Account No. 09-0468 in the amount of $ 710.00 .

x The Commissioner is hereby authorized to charge payment of the following fees associated with this communication or credit any overpayment to Deposit Account No. 09-0468. A duplicate copy of this sheet is enclosed.

x Any additional filing fees required under 37 CFR 1.16.

x Any patent application processing fees under 35 CFR 1.17.

Respectfully submitted,

By Louis J. Percello

Louis J. Percello
Registration No.: 33,206
Tel. (914) 945-3145

IBM CORPORATION
INTELLECTUAL PROPERTY LAW DEPT.
P.O. BOX 218
YORKTOWN HEIGHTS, NY 10598

Express Mail EL559661095US
Date of Deposit: Nov. 15, 2000

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re application of: Eric W. Brown et al.
Serial No.:
Filed: Herewith
For: FINDING THE MOST LIKELY ANSWER TO A NATURAL LANGUAGE QUESTION

Group No.
Examiner:

Assistant Commissioner of Patents and Trademarks
Washington, D.C. 20231

EXPRESS MAIL CERTIFICATE

Express Mail Label Number EL559661095US1
Date of Deposit November 15, 2000

I hereby certify that the attached paper or fee

Patent Application Transmittal Letter (original and one copy)
Patent Application
Unsigned Declaration and Power of Attorney
Informal Drawings (9 Sheets) NINE
Return Postcard

is being deposited with the United States Postal Service "Express Mail Post Office to Addressee" service under 37 CFR 1.10 on the date indicated above and is addressed to the Assistant Commissioner of Patents and Trademarks, Washington, D.C. 20231.

Kathy Cognatello
(Name)

Kathy Cognatello
(Signature)

Note: Each paper must have its own certificate and the "Express Mail" label number as a part thereof or attached thereto. When, as here, the certification is presented on a separate sheet, that sheet must (1) be signed and (2) fully identify and be securely attached to the paper or fee it accompanies. Identification should include the serial number and filing date of the application as well as the type of paper being filed, e.g. complete application, specification and drawings, responses to rejection or refusal, notice of appeal, etc. If the serial number of the application is not known, the identification should include at least the name of the inventor(s) and the title of the invention.

Note: The label number need not be placed on each page. It should, however, be placed on the first page of each separate document, such as, a new application, amendment, assignment, and transmittal letter for a fee, along with the certificate of mailing by "Express Mail". Although the label number may be on checks, such a practice is not required. In order not to deface formal drawings it is suggested that the label number be placed on the back of each formal drawing or the drawings be accompanied by a set of informal drawings on which the label number is placed.

Docket No. YOR920000807US1

SYSTEM AND METHOD FOR FINDING THE MOST LIKELY ANSWER TO A NATURAL LANGUAGE QUESTION

CROSS-REFERENCE TO A RELATED PATENT APPLICATION

This patent application is related to commonly-assigned U.S. Patent Application S.N. 09/495645 (IBM docket YOR919990503US2) filed February 1, 2000, by Brown et al., entitled "System, Method, and Program Product for Answering Questions using a Search Engine", the disclosure of which is incorporated by reference herein in its entirety.

FIELD OF THE INVENTION

This invention relates to the field of automated question answering. More specifically, the invention relates to the selection of an answer to a question from a pool of potential answers which are manually or automatically extracted from a large collection of textual documents.

BACKGROUND OF THE INVENTION

Information retrieval (IR) is the process of locating documents in a collection or from an unbounded set such as the Web based on an expression of a human user's information need. The user's information need is typically expressed in the form of a query which consists of a set of keywords and/or logical operators. A particular type of information retrieval is Question Answering (Q&A).

Unlike information retrieval, in Q&A the user expresses his or her information need in the form of a factual natural language question (e.g., "who played Don Vito Corleone in the movie 'The Godfather'?").

Unlike information retrieval, Q&A returns a short snippet or snippets of text (e.g., phrases) which provide the exact answer to the question rather than a document or set of documents related to the question.

Unlike information retrieval, Q&A systems must understand the user's questions to a deeper level, e.g., properly dealing with negations ("Not") and/or the question's discourse, logical, or temporal structure ("Which U.S. president succeeded Nixon?", "What is the smallest country in Europe?"). When given an input such as "What is the capital of India?", an IR-based

system will typically return documents about India and about capital (in all of its possible senses) but not necessarily documents which contain the answer "New Delhi". Q&A systems will weed out the wrong senses of capital (e.g., "financial capital") and concentrate on the meaning of that word ("head city") that fits best the overall structure of the question.

Example:

	Information Retrieval (IR)	**Question Answering (Q&A)**
user input	query: environmental protection organizations in developing countries	natural language question: What is the capital of Delaware?
system output	documents: http://www.envorg.br http://www.ioe.int	answer(s): Dover
input processing	keyword based	natural language based
main advantage	can return partially related documents	precise answer, not buried in unneeded text

Further information on information retrieval and text analysis can be found in, for example, Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval", ACM Press, New York, 1999; Ravin and Wacholder, "Extracting Names from Natural-Language Text", IBM Research Report 20338, 1996; Byrd and Ravin, "Identifying and Extracting Relations in Text", Proceedings of NLDB 99, Klagenfurt, Austria, 1999. Further information on Question Answering can be found in Kupiec, "MURAX: A Robust Linguistic Approach For Question Answering Using An On-Line Encyclopedia", *Proc. of SIGIR 1993*, Pittsburgh, PA, 1993; Prager et al., "The Use of Predictive Annotation for Question-Answering in TREC8", *Proc. of TREC8*, Gaithersburg, MD, 2000; Prager, "Question-Answering by Predictive Annotation", *Proc. of SIGIR 2000*, Athens, Greece, 2000; Radev et al., "Ranking Suspected Answers to Natural Language Questions using Predictive Annotation", *Proc. of ANLP'00*, Seattle, WA, 2000.

STATEMENT OF PROBLEMS WITH THE PRIOR ART

Recently, some search engines accessible from the Web have started to provide question answering services. A notable example is Ask Jeeves (www.ask.com) (Ask Jeeves and Ask.com are service marks of Ask Jeeves, Inc.). Ask Jeeves uses a fairly simple keyword-based approach to

give the user a feeling of a "natural language interface". For example, a question such as "What is the capital of Kenya" is apparently correctly interpreted but it returns pointers to several Web sites with information about Kenya, one of which does include the correct answer ("Nairobi"). However, related questions such as "How long does it take to fly from New York to London on the Concorde" produces instead a set of questions related to the original question asked by the user. The user then has to select which of the suggested paraphrases is most likely to return answers to the original question. Examples of such follow-up questions include "Where can I find cheap flights from the UK?". The method used to produce answers apparently consists of five steps: (a) partially parse the query; (b) map the query to a canned set of manually produced questions or question templates; (c) map canned questions to existing knowledge bases (Ask Jeeves points to other people's web sites for the real data: FAQs, authoritative pages, travel agencies, etc.); (d) do a meta search on 5 big search engines (and return their answers); and (e) if there is no match in "b" then record the query for later human analysis.. Note that "b" is essentially a person-intensive task – it involves the creation of a list of key phrases and the canned questions that they map to (and then the underlying pages that they map to).

Two things that systems such as Ask Jeeves don't do are: (a) provide a precise answer to a factual question; and (b) restrict their output to only the relevant answer by getting rid of other text from the same document that does not include the answer. A third problem with such systems is their overly large dependence on human knowledge engineering.

OBJECTS OF THE INVENTION

An object of this invention is an improved system, method, and program product for answering natural language questions from either network sites or from document collections physically or virtually residing on the user's local area network (LAN) or intranet.

An object of this invention is an improved system, method, and program product for providing precise answers to factual questions.

An object of this invention is an improved system, method, and program product which outputs an answer to a user's question without adding unnecessary content around the answer.

An object of this invention is an improved system, method, and program product which contains an embodiment of a natural language component that better analyzes and understands

queries asked in the form of questions.

An object of this invention is an improved system, method, and program product which uses a mathematical model of properties of textual documents to provide better understanding of the user's question and a better set of resulting answers.

SUMMARY OF THE INVENTION

The foregoing and other problems are overcome by methods and apparatus in accordance with embodiments of this invention.

This invention is a computer system, method, and program product that contains a feature extraction module, a feature combination module, an answer selection module, and an answer presentation module.

The feature extraction module computes automatically certain properties of the question and the documents from which the answer is to be extracted. Among these properties, potential answers to the question are also extracted and annotated with the features already extracted.

The feature combination module provides an automated mechanism for characterizing properties of the documents and question as well as the features and potential answers extracted by the feature extraction module.

The answer selection module ranks the potential answers based on an objective score produced by the feature combination module. As a result, answers that are more likely to represent the correct answer to a question are ranked higher.

The answer presentation module presents the highest ranked potential answers to the user by providing a variable (specified by the user) amount of context.

BRIEF DESCRIPTION OF THE DRAWINGS

The above set forth and other features of the invention are made more apparent in the ensuing Detailed Description of the Invention when read in conjunction with the attached Drawings, wherein:

Figure 1 is an overall block diagram of the basic architecture of the invention.

Figure 2 depicts an example of questions posed by the user.

Figure 3 describes the expected input by the invention in the form of annotated (or

indexed) document passage or passages.

Figure 4 illustrates some sample features (7) extracted by the feature extraction module (1) as well as the output (8) of the feature combination module (2) shown in the TOTAL column and the list of potential answers (9).

Figure 5 indicates which potential answers (11) from the set (10) have been selected by the answer selection module (3).

Figure 6 displays the output of the answer presentation module (4). The output can consist of either (a) the top-ranked answers or (b) the top-ranked answers plus some context, or (c) the documents in which the top-ranked answers occur. In all cases, a pointer to the original document may be included.

Figure 7 is a flowchart showing how the EXECUTION component of the invention operates.

Figure 8 is a flowchart of the TRAINING component of the invention.

Figure 9 is a flowchart showing the "EXTRACT FEATURES" (805) and "COMPUTE COMPOSITE SCORE" (806) procedures which are jointly used in the EXECUTION and the TRAINING component.

DETAILED DESCRIPTION OF THE INVENTION

The present invention better satisfies a user's information need by providing precise answers to factual natural language questions.

Figure 1 shows the basic architecture of the system in a non-limiting preferred embodiment. The system contains at least four components: a feature extraction module (1), a feature combination module (2), an answer selection module (3), and an answer presentation module (4).

An indexed set of document passages (6) is suspected to contain an answer to the user's natural language question (5). The feature extraction module (1) computes a set of features from the input documents (6) and the natural language question (5). These features are stored in per-document, per-question feature set (7). Among the features that are used in (1) are (a) the proximity of words from the question to words from the document; (b) the number of overlapping words between the question and the document, (c) the number of times that a given document

contains the same text, etc.

An example of a user question (5) is shown in Figure 2. Item 201 is the user question. Similarly, an example of the indexed documents in the preferred embodiment appears in Figure 3. Item 300 represents a sample annotated input passage. In the preferred embodiment and as illustrated in Figure 3, the input passage consists of several parts: an index to the document that contains the passage (301), an optional passage score (302), an annotated representation, or processed query (303) of the user question (5), and an annotated representation (304) of the passage (6).

The different features in the feature set are combined by the feature combination module (2, shown also on Figure 9) to provide a set of composite features (8), one or more per question-document pair. Based on the feature set (7), the feature combination module also produces a set of potential answers (9) which contains the likely answers to the question (5) as well as zero or more less likely answers. In the preferred embodiment, the following features (402—409) are used, among others. Figure 9 is a flowchart describing the process of computing the different features and combining them into a single score.

Feature 402 (type) is the semantic type of the current suspected answer. For example, the semantic type of "Lou Vasquez" is "PERSON". The processed query (303) indicates the semantic type of the potential answers that are most likely to be relevant to the given question. For example, the semantic types of the most likely answers to a Who question are "PERSON", "ORG", "NAME", and "ROLE" as indicated in 303.

Feature 403 (number) represents the position of the suspected answer among all suspected answers within all document passages. Example: "Lou Vasquez" is the first suspected answer in 400.

Feature 404 (rspanno) is the position of the suspected answer among all suspected answers within the given passage. Example: "Derric Evans" is the fourth suspected answer within the passage in which it appears.

Feature 405 (count) is the number of suspected answers of a given semantic type retrieved within a given passage.

Feature 406 (notinq) represents the number of words in a suspected answer that do not appear in the user question. Example: Notinq ("Woodbridge high school") = 1, because both

"high" and "school" appear in the query while "Woodbridge" does not. Whenever the actual value of notinq is zero, then the value is replaced with a very high negative value to indicate that the current potential answer is highly unlikely to be correct.

Feature 407 (type) is the position of the semantic type in the list of potential semantic types for the question. Example: Type ("Lou Vasquez") = 1, because the span type of "Lou Vasquez", namely "PERSON" appears first in the list of potential span types, "PERSON ORG NAME ROLE".

Feature 408 (avgdst) represents the average distance in words between the beginning of the potential answer and the words in the question that also appear in the passage. Example: given the passage "Tim O'Donohue, Woodbridge High School's varsity baseball coach, resigned Monday and will be replaced by assistant Johnny Ceballos, Athletic Director Dave Cowen said." and the span "Tim O'Donohue", the value of avgdst is equal to 8.

Feature 409 (Sscore) is the passage relevance as computed the retrieval engine.

Other features that are not included in the example here include (a) the frequency of a given potential answer on the list, (b) the semantic relation between words from the question and words from the potential answer, and (c) a strength score that is a function of the relevance score 409.

The feature combination module (see item 710 in Figures 7 and 8) uses either a statistical or a manual technique to determine the best formula to combine the different features. A statistical technique used in the preferred embodiment is logistic regression (prior art). In the preferred embodiment, these techniques work as follows: a set of questions and lists of potential answers are annotated semantically. All features are computed and based on developer-specified correct answers, the system learns (item 808, see Figure 8) the proper function to combine all features. In alternative embodiments, the function can be specified manually. In the preferred embodiment, the function is a linear combination of all features:

$$F = \sum_{i=1}^{n} w_i f_i$$

In this formula, F is the composite function; f_i are the different features used, w_i are the weights associated with these features, and n is the total number of features.

The learning system 808 operates as follows: a set of questions if obtained through one or more interfaces (802), the matching passages (803) are obtained using a search engine or by other means, including manually, potential answer passages (804) are extracted, features (805) are also extracted (as in Figure 7), The composite score is computed (806), then all potential answers are ranked based on their score (807), and the ranking, along with the set of features is used to learn a discrimination function (808) which will be later used to classify potential answers into more or less likely ones.

The answer selection module (3) uses the composite features (8) and the set of potential answers (9) to produce a ranked list of potential answers (10). Answers near the beginning of that ranked list are assumed to be more likely answers to the original question (5). In more detail, the answer selection module is described in Figure 7. The first step is to get a question (702), then decide whether it is of the correct type (703, factual question, e.g., but not limited to why, where, and how much questions). If the question is not of the right type, the system rejects it and defaults on a search engine (704). Otherwise, the system extracts the type of the question (including, but not limited to when, what, where questions). Next, a search engine (at least, in the preferred embodiment) is used (706) to extract matching passages to the query. If there are no matching passages (707), the system says so (708). Otherwise, control is passed to box 709 which decides whether the documents containing the potential answers returned by the search engine do contain the answer at all, to begin with. If no, the system again falls back on IR (715). If there are answers, the system extracts their type (e.g., why, where, etc.), see box 710. The next two boxes: 805 and 806 are the same as during the training stage (Figure 8). The final box (713) selects the highest ranking documents,

In Figure 4, the likeliest answers to the user question are the ones with the highest composite score. Example: "Lou Vasquez" has a score of –9.93 which is higher than all other scores. Figure 5 shows an example containing a number of the highest-ranking potential answers (501).

Finally, the answer presentation module (4) adds a certain amount of context (zero or more characters) to the answers ranked highest within the ranked list (10). The resulting set of answers along with the neighboring contest are presented to the user in the form of a set of likeliest answers (11). Figure 6 shows an example (600) from the preferred embodiment which

indicates all of the following: (a) the highest-ranking answers (601), (b) their scores (602), (c) the context in which they appear (603), and (d) pointers to the document where they appear (604).

The answer presentation ascertains that the same potential answer doesn't appear more than once in context. If the context is too large and allows for more than one highly-ranked potential answer to be included, the answer presentation module inserts additional, lower-ranked answers to make use of all space available.

Operation of the invention

In the preferred embodiment, the invention can handle arbitrary factual questions, including but not limited to where-, when-, what-, how-, and what- questions. The set of documents from which the answers are extracted can be one of the following: a collection of documents physically or virtually residing on the user's local area network (LAN) or intranet, an indexed encyclopedia, or the entire Web, or any combination of the above. The user can specify one or more questions using one or more interfaces. The invention analyzes the question or questions as well as the entire collection of documents.

Example:

User inputs a question "Who was Johnny Mathis' high school track coach?" in box (5) of Figure 1. The input to the system consists of the user input (5) as well as a set of document passages (6) that are deemed likely to contain answers to the user question (5). In the preferred embodiment, the related passages are retrieved by an information retrieval system (or search engine) which may be similar to the one described in Patent Filing IBM Y0999-503. The structure of the user question (5) is shown in Figure 2 while the document passage input (6) is presented in Figure 3.

The invention analyzes the user question and stores all words from the question, as well as a representation of the logical structure of the question. For the document collection, the invention extracts portions that are considered most likely to contain a factual answer related to the user question. These portions are analyzed and annotated with a number of automatically computed features, similar to but not limited to the ones shown in Table 1. The composite feature is shown in the TOTAL column of Table 1.

The contents of Table 1 are sorted based on the composite feature. The highest ranked answers (as shown in the first column) are presented to the user, possibly in a user-specified context. In the example, the likeliest answer to the question "Who was Johnny Mathis' high school track coach" is "Lou Vasquez" with a total score of –9.93 which is higher than all other scores.

CLAIMS

We claim:

1. A method for selecting answers to natural language questions from a collection of textual documents comprising the steps of

extracting scoring features from a candidate list of passages of possible answers;

scoring the possible answers using the extracted features and a feature scoring function; and

presenting the best scoring possible answer to the user with context from the passage containing the answer.

2. A method as in claim 1, wherein the features used to score possible answers consists of one or more of the following features: a semantic type of a current suspected answer, a position of the suspected answer among all suspected answers within all document passages, a position of the suspected answer among all suspected answers within the given passage, a number of suspected answers of a given semantic type retrieved within a given passage, a number of words in a suspected answer that do not appear in the user question, a position of the semantic type in the list of potential semantic types for the question, an average distance in words between the beginning of the potential answer and the words in the question that also appear in the passage, a passage relevance as computed by the information retrieval engine, a frequency of a given potential answer on the list, a semantic relation between words from the question and words from the potential answer, and a strength score that is a function of the relevance score.

3. A method as in claim 2, wherein the feature scoring function is a linear combination of weighted features.

4. A method as in claim 3, wherein the parameters of the scoring function are manually determined.

5. A method as in claim 3, wherein the parameters of the scoring function are learned by a machine learning algorithm.

6. A method as in claim 1 where the candidate list of passages of possible answers is obtained from the collection of documents using an information retrieval engine

7. A computer system that extracts answers to natural language questions from a large collection of textual documents consisting of one or more of the following modules:

a feature extraction module;

a feature combination module, containing a "feature extraction" and "compute composite score" components;

an answer selection module; and

an answer presentation module.

8. A computer system, as in claim 7, wherein the feature extraction module extracts one or more of the following features: a semantic type of the current suspected answer, a position of the suspected answer among all suspected answers within all document passages, a position of the suspected answer among all suspected answers within the given passage, a number of suspected answers of a given semantic type retrieved within a given passage, a number of words in a suspected answer that do not appear in the user question, a position of the semantic type in the list of potential semantic types for the question, an average distance in words between the beginning of the potential answer and the words in the question that also appear in the passage, a passage relevance as computed the retrieval engine, a frequency of a given potential answer on the list, a semantic relation between words from the question and words from the potential answer, and a strength score that is a function of the relevance score.

9. A computer system as in claim 7, wherein the feature combination module employs a feature scoring function with parameters learned by a machine learning method.

10. A computer system as in claim 7, wherein the answer selection module selects the answer with the best score obtained from the feature combination module.

11. A computer system as in claim 7, wherein the answer presentation module shows the top scored answer within the context as specified by a user or a system.

12. A computer program product that performs the steps of :

determining a feature scoring function during a training phase either manually or via a machine learning algorithm applied to a set of training questions, corresponding answer passages, and certain extracted features; and

during an execution phase, extracting certain features from questions and corresponding possible answer phrases, applying the feature scoring function determined during the training phase to score each possible answer phrase, selecting one or more of the best scoring answer phrases, and displaying the answer phrases to the user with optional additional context from the answer passages.

SYSTEM AND METHOD FOR FINDING THE MOST LIKELY ANSWER TO A NATURAL LANGUAGE QUESTION

ABSTRACT

Automated question answering is disclosed that relates to the selection of an answer to a question from a pool of potential answers which are manually or automatically extracted from a large collection of textual documents. The a feature extraction component, a feature combination component, an answer selection component, and an answer presentation component, among others, are included. The input to the system is a set of one or more natural language questions and a collection of textual document. The output is a (possibly ranked) set of factual answers to the questions, these answers being extracted from the document collection.

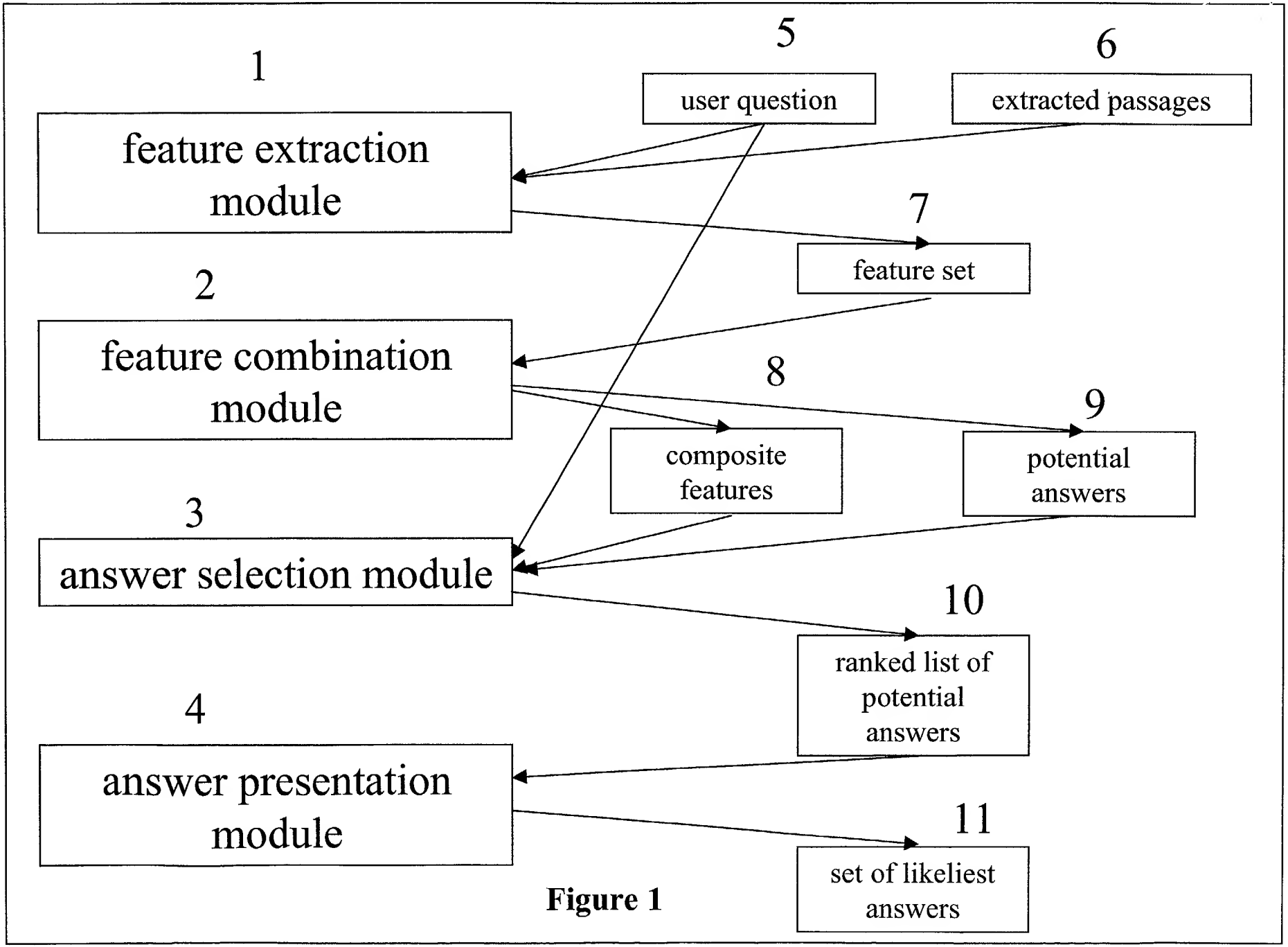
1
feature extraction module
2
feature combination module
3
answer selection module
4
answer presentation module
5
user question
6
extracted passages
7
feature set
8
composite features
9
potential answers
10
ranked list of potential answers
11
set of likeliest answers

Figure 1

201

Who was Johnny Mathis' high school track coach?

Figure 2

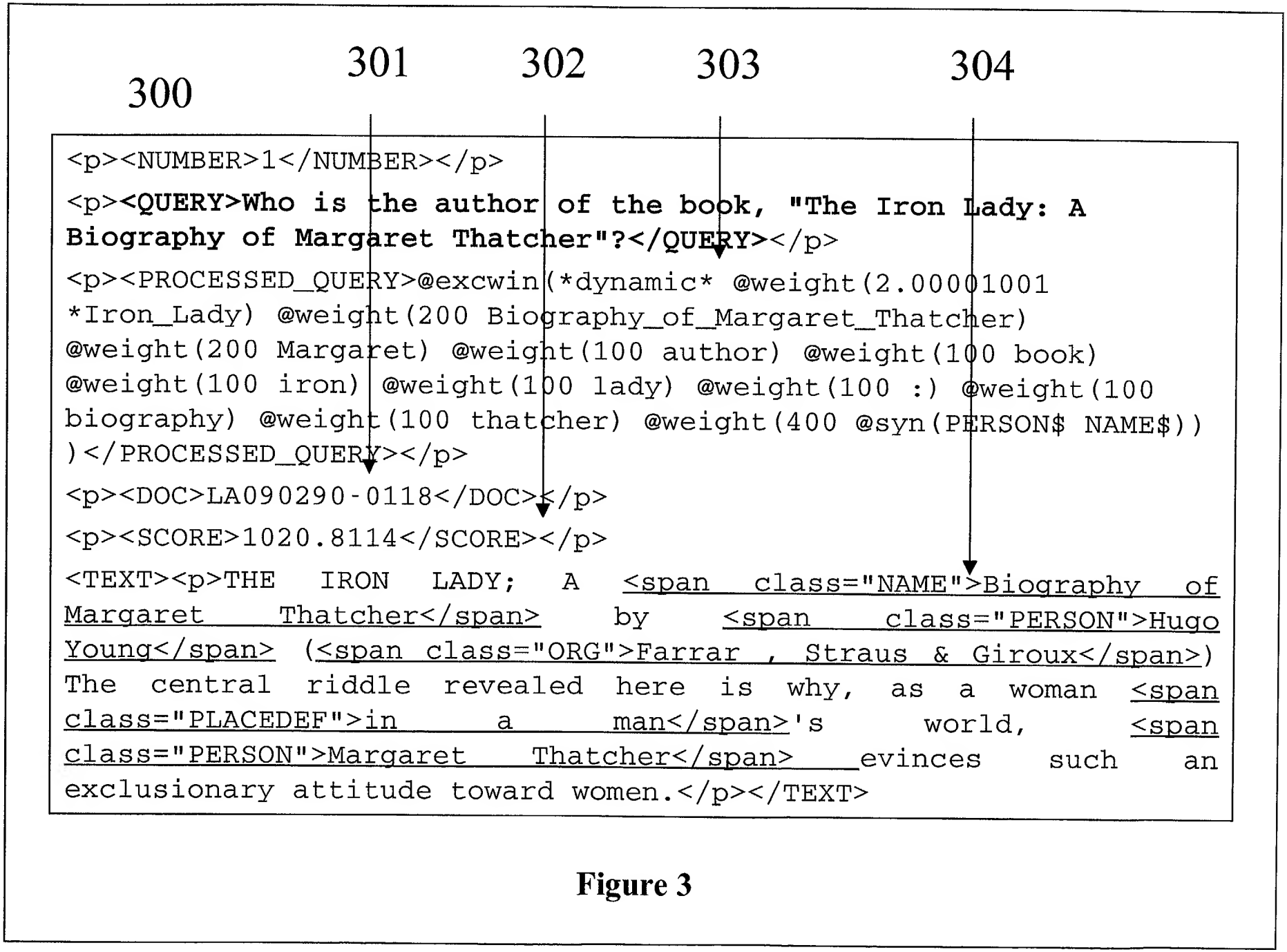
300
301
302
303
304
<NUMBER>1</NUMBER>
<QUERY>Who is the author of the book, "The Iron Lady: A Biography of Margaret Thatcher"?</QUERY>
<PROCESSED_QUERY>@excwin(*dynamic* @weight(2.00001001 *Iron_Lady) @weight(200 Biography_of_Margaret_Thatcher) @weight(200 Margaret) @weight(100 author) @weight(100 book) @weight(100 iron) @weight(100 lady) @weight(100 :) @weight(100 biography) @weight(100 thatcher) @weight(400 @syn(PERSON$ NAME$)))</PROCESSED_QUERY>
<DOC>LA090290-0118</DOC>
<SCORE>1020.8114</SCORE>
<TEXT> THE IRON LADY; A Biography of Margaret Thatcher by Hugo Young (Farrar , Straus & Giroux) The central riddle revealed here is why, as a woman in a man's world, Margaret Thatcher evinces such an exclusionary attitude toward women. </TEXT>

Figure 3

400

401 Potential answer	402 Type	403 Number	404 Rspanno	405 Count	406 Notinq	407 Type	408 Avgdst	409 Sscore	410 TOTAL
Lou Vasquez	PERSON	1	1	6	2	1	16	0.02507	-9.93
Tim O'Donohue	PERSON	17	1	4	2	1	8	0.02257	-12.57
Athletic Director Dave Cowen	PERSON	23	6	4	4	1	11	0.02257	-15.87
Johnny Ceballos	PERSON	22	5	4	1	1	9	0.02257	-19.07
Civic Center Director Martin Durham	PERSON	13	1	2	5	1	16	0.02505	-19.36
Johnny Hodges	PERSON	25	2	4	1	1	15	0.02256	-25.22
Derric Evans	PERSON	33	4	4	2	1	14	0.02256	-25.37
NEWSWIRE Johnny Majors	PERSON	30	1	4	2	1	17	0.02256	-25.47
Woodbridge High School	ORG	18	2	4	1	2	6	0.02257	-28.37
Evan	PERSON	37	6	4	1	1	14	0.02256	-29.57
Gary Edwards	PERSON	38	7	4	2	1	17	0.02256	-30.87
O.J. Simpson	NAME	2	2	6	2	3	12	0.02507	-37.40
South Lake Tahoe	NAME	7	5	6	3	3	14	0.02507	-40.06
Washington High	NAME	10	6	6	1	3	18	0.02507	-49.80
Morgan	NAME	26	3	4	1	3	12	0.02256	-52.52
Tennesseefootball	NAME	31	2	4	1	3	15	0.02256	-56.27
Ellington	NAME	24	1	4	1	3	20	0.02256	-59.42
assistant	ROLE	21	4	4	1	4	8	0.02257	-62.77
the Volunteers	ROLE	34	5	4	2	4	14	0.02256	-71.17
Johnny Mathis	PERSON	4	4	6	-100	1	11	0.02507	-211.33
Mathis	NAME	14	2	2	-100	3	10	0.02505	-254.16
coach	ROLE	19	3	4	-100	4	4	0.02257	-259.67

Figure 4

501

Lou Vasquez
Tim O'Donohue
Athletic Director Dave Cowen
Johnny Ceballos
Civic Center Director Martin
Durham
Johnny Hodges
Derric Evans

Figure 5

600

Document ID	Score	Extract
LA053189-0069	-9.93	Lou Vasquez , track coach of O.J. Simpson , Ollie Matson and Johnny Mathis during his 32-year career, died Saturday while at his South Lake Tahoe vacation cabin, it was announced Tuesday . He was 68 . His Washington High school teams won five consecu
LA060889-0181	-12.57	Tim O'Donohue , Woodbridge High School 's varsity baseball coach , resigned Monday and will be replaced by assistant Johnny Ceballos , Athletic Director Dave Cowen said.
LA062090-0017	-19.36	Civic Center Director Martin Durham said Mathis was to have entered the parking lot in a convertible Rolls-Royce to cut the ribbon for the dedication of Johnny Mathis [illegible]
LA052390-0122	-25.22	Ellington liked what he heard. Johnny Hodges was quitting the band and Morgan was invited to replace him, but he couldn't leave high school to go on the road. The new album's title track and " In a Sentimental Mood " are Morgan 's most recent homages
LA062389-0083	-25.37	NEWSWIRE Johnny Majors , Tennessee football coach , said that prize recruit Derric Evans will not be allowed to play for the Volunteers because of his arrest Tuesday night in Dallas . Evans and a high school teammate, Gary Edwards , were charged with

604 602 603 601

Figure 6

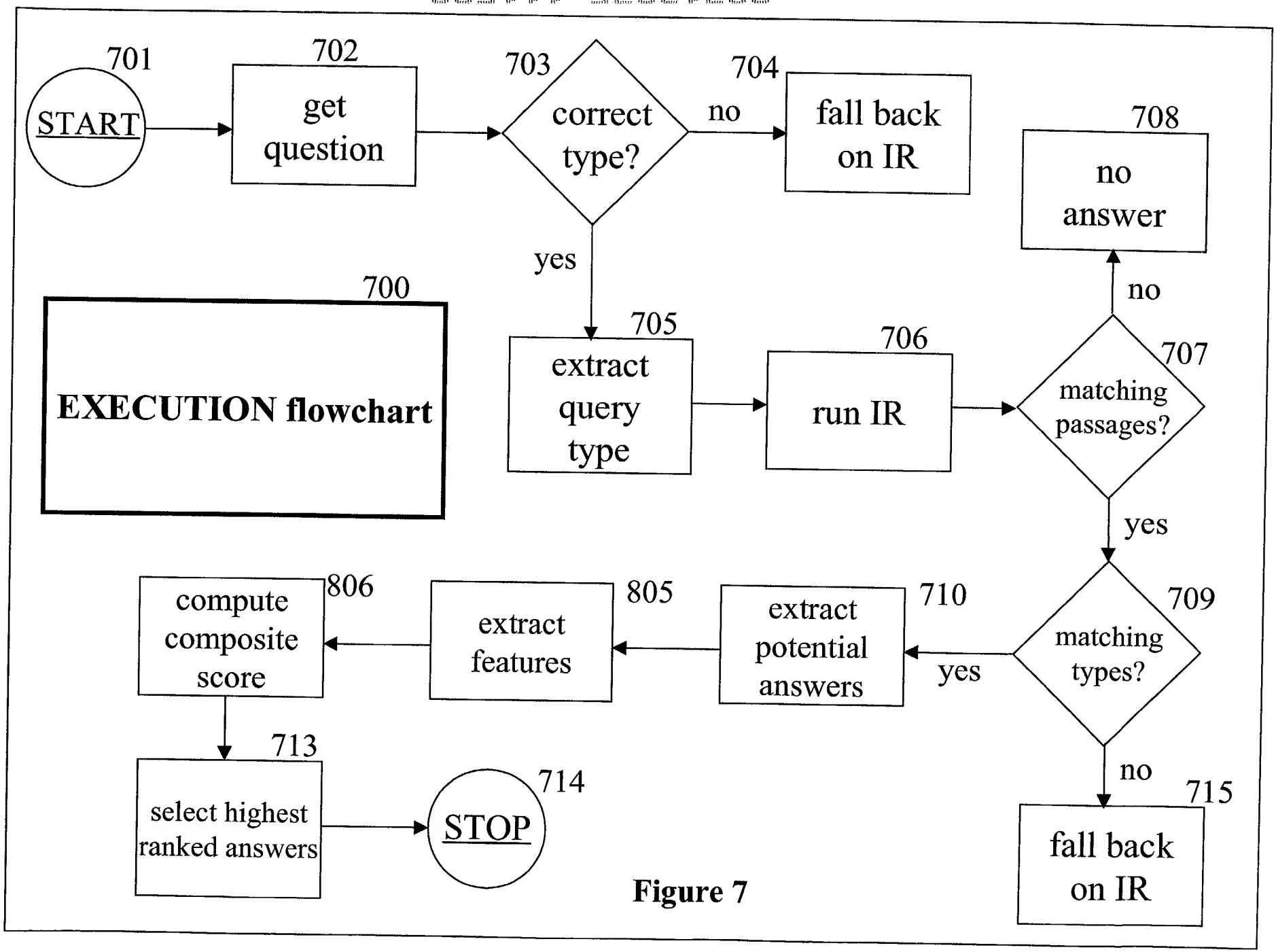
701
START
702
get question
703
correct type?
no
704
fall back on IR
yes
700
EXECUTION flowchart
705
extract query type
706
run IR
707
matching passages?
no
708
no answer
yes
709
matching types?
yes
710
extract potential answers
805
extract features
806
compute composite score
713
select highest ranked answers
714
STOP
no
715
fall back on IR

Figure 7

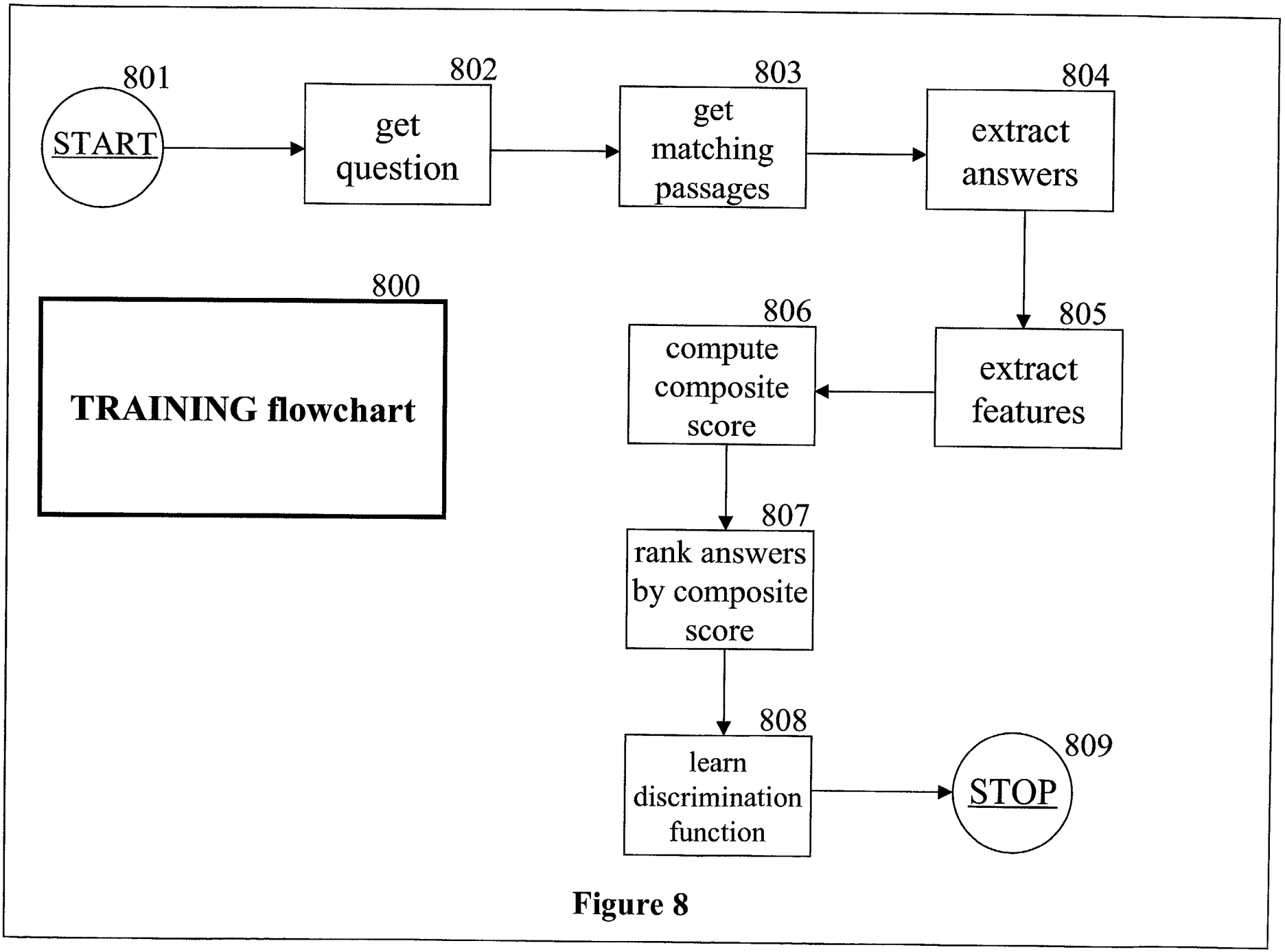
801
START
802
get question
803
get matching passages
804
extract answers
800
TRAINING flowchart
806
compute composite score
805
extract features
807
rank answers by composite score
808
learn discrimination function
809
STOP

Figure 8

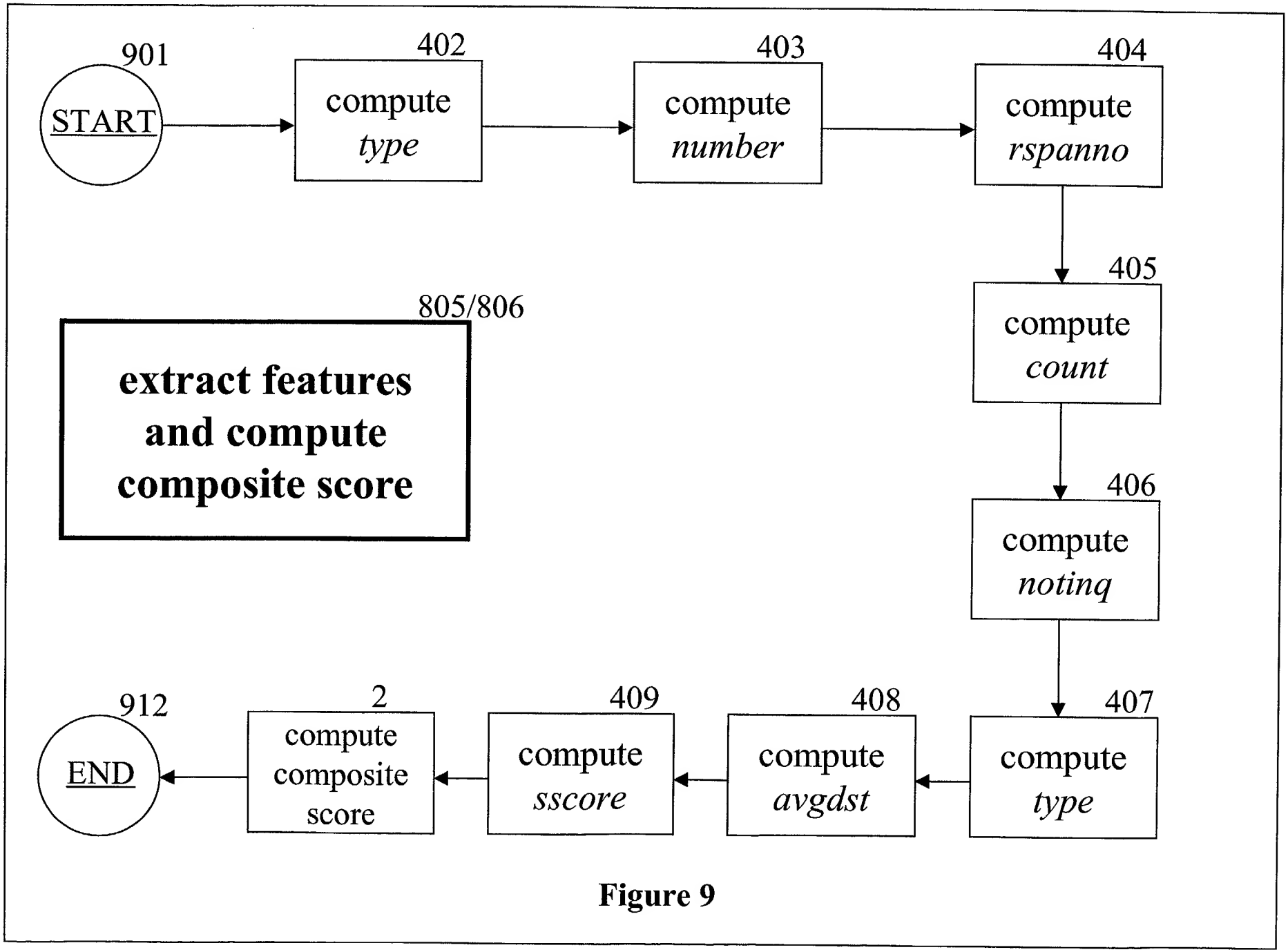
901
START
402
compute
type
403
compute
number
404
compute
rspanno
405
compute
count
406
compute
notinq
407
compute
type
408
compute
avgdst
409
compute
sscore
2
compute
composite
score
912
END
805/806
extract features
and compute
composite score

Figure 9

DECLARATION AND POWER OF ATTORNEY FOR PATENT APPLICATION

As a below named inventor, I hereby declare that:

My residence, post office address and citizenship are as stated below next to my name;

I believe I am the original, first and sole inventor (if only one name is listed below) or an original, first and joint inventor (if plural names are listed below) of the subject matter which is claimed and for which a patent is sought on the invention entitled:

SYSTEM AND METHOD FOR FINDING THE MOST LIKELY ANSWER TO A NATURAL LANGUAGE QUESTION

The specification of which (check one)

X is attached hereto.

____ was filed on _____ as United States Application Number

or PCT International Application Number _____

and was amended on _____ (if applicable)

I hereby state that I have reviewed and understand the contents of the above identified specification, including the claims, as amended by any amendment referred to above.

I acknowledge the duty to disclose information which is material to the patentability of this application in accordance with Title 37, Code of Federal Regulations, Section 1.56.

I hereby claim foreign priority benefits under Title 35, United States Code, §119(a)-(d) or §365(b) of any foreign application(s) for patent or inventor's certificate, or §365(a) of any PCT International application which designated at least one country other than the United States, listed below and have also identified below, by checking the box, any foreign application for patent or inventor's certificate, or PCT International application, having a filing date before that of the application on which priority is claimed:

Prior Foreign Application(s)			Priority Claimed
_____ (Number)	_____ (Country)	_____ (Day/Month/Year Filed)	___ Yes ___ No
_____ (Number)	_____ (Country)	_____ (Day/Month/Year Filed)	___ Yes ___ No
_____ (Number)	_____ (Country)	_____ (Day/Month/Year Filed)	___ Yes ___ No

I hereby claim the benefit under 35 U.S.C. §119(e) of any United States provisional application(s) listed below.

_____ (Application Number)	_____ (Filing Date)
_____ (Application Number)	_____ (Filing Date)

I hereby claim the benefit under 35 U.S.C. §120 of any United States Application(s), or §365(c) of any PCT International application designating the United States, listed below and, insofar as the subject matter of each of the claims of this application is not disclosed in the prior United States, or PCT International application in the manner provided by the first paragraph of 35 U.S.C. §112, I acknowledge the duty to disclose information material to the patentability of this application as defined in 37 CFR §1.56 which occurred between the filing date of the prior application and the national or PCT international filing date of this application:

_____ (Application Serial No.)	_____ (Filing Date)	_____ (Status) (patented, pending, abandoned)
_____ (Application Serial No.)	_____ (Filing Date)	_____ (Status) (patented, pending, abandoned)

I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and that willful false statements may jeopardize the validity of the application or any patent issued thereon.

POWER OF ATTORNEY: As a named inventor I hereby appoint the following attorney(s) and/or agent(s) to prosecute this application and transact all business in the Patent and Trademark Office connected therewith (list name and registration number).

Manny W. Schecter (Reg. 31,722), Lauren Bruzzone (35,082), Christopher A. Hughes(Reg. 26,914), Edward A. Pennington (Reg. 32,588), John E. Hoel (Reg. 26,279), Joseph C. Redmond, Jr. (Reg. 18,753), Douglas W. Cameron (Reg. 31,596), Louis P. Herzberg (Reg. 41,500), Derek S. Jennings (Reg. 41,473), Stephen C. Kaufman (Reg. 29,551), Daniel P. Morris (Reg. 32,053), Paul J. Otterstedt (Reg. 37,411), Louis J. Percello (Reg. 33,206), Robert P. Tassinari, Jr. (36,030), Robert M. Trepp (Reg. 25,933) and Marian Underweiser (Reg. 46,134)

Send Correspondence to: Louis J. Percello, Intellectual Property Law Dept.

IBM Corporation, P.O. Box 218, Yorktown Heights, New York 10598

Direct Telephone Calls to: (name and telephone number) Louis J. Percello (914)945-3145

Eric William Brown
Full name of sole or first inventor

_____
Inventor's Signature Date

13 Indian Hill Road, New Fairfield, Connecticut 06812-2544
Residence

USA
Citizenship

same as above
Post Office Address

Express Mail EL559661095US
Date of Deposit: Nov. 15, 2000

DECLARATION AND POWER OF ATTORNEY FOR PATENT APPLICATION

Anni R. Coden
Full name of second joint-inventor, if any

Inventor's signature Date

750 Kapock Street, Apt. 1403, Bronx, New York 10463
Residence

USA
Citizenship

same as above
Post Office Address

John Martin Prager
Full name of third joint-inventor, if any

Inventor's signature Date

3 Lochsley Lane, Pomona, New York 10970
Residence

USA
Citizenship

same as above
Post Office Address

Dragomir Radkov Radev
Full name of fourth joint-inventor, if any

Inventor's Signature Date

3296 Columbus Lane, Ann Arbor, Michigan 48103
Residence

Bulgaria
Citizenship

same as above
Post Office Address

Valerie Samn
Full name of fifth joint inventor, if any

Inventor's Signature Date

3133 Broadway, Apt. 7, New York, New York 10027
Residence

USA
Citizenship

same as above
Post Office Address

Full name of sixth joint-inventor, if any

Inventor's signature Date

Residence

Citizenship

Post Office Address